Case 3:23-cv-03417-VC Document 504-2 Filed 03/28/25 Page 1 of 6

EXHIBIT A

LLM Datasets Considerations 3/31 Meta Confidential A/C PRIV

Goals: (from Ahmad)

- 1) Get as much long form writing as possible in the next 4-6 weeks
 - Get Instant Articles that have already been licensed to Meta (check Meta contract); same could be for video content uploaded
 - Books all genres
 - Movie Scripts or transcripts via ASR
 - Magazines
- Speed up International launches with crowd sourced human raters RLHF (Reinforcement Learning w Human Feedback) with some rewards. Ex: launch in India.
- 3) Launch in Q3 and get the data within the coming weeks

Want to address goal #1 below and Redacted - Privilege

A: Datasets examples that will benefit the eng team if possible to use:

- 1. <u>Libgen</u> (fictions, non-fictions, journals, magazines)
- 2. Simplyscripts (scripts)
- 3. Internetmoviescript (scripts)
- 4. https://manybooks.net/ (free books)
- 5. https://archiveofourown.org/ (fan work)
- 6. https://www.fanfiction.net/ (fan fiction)
- https://www.fictionpress.com/ (fan fiction)
- 8

Redacted - Privilege



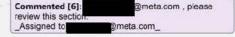
Redacted - Privilege

Redacted - Privilege

Appendix:

Info gathered from initial outreach and back channels from the Media Partnership team so far.

- Unclear that all publishers have legal rights to license books for Al training. If not, they will need to go back to the content owners to get permission which will take much longer. Please note that Al BD team has had discussions with text book publishers and some have confirmed that they do have the rights to license.
- 2. Have not seen OpenAI securing any rights for AI training even through back channels with publishers. Meta will have to do the work to become the first company to do these license deals.



Document Revisions Total Revisions: 1

Author: Logan Kerr

Date: 4/6/2023 6:52:00 PM

Type: Delete

Range:

Document Comments Total Comments: 6 Author: Alex Yu Date: 4/1/2023 12:40:00 AM Range: I am still collecting from the eng team. I may add more later. @meta.com please add in the doc or comment. thanks! Scope: Datasets examples that will benefit the eng team if possible to use: Author: Angela Fan Date: 4/1/2023 2:08:00 AM Range: https://archiveofourown.org/https://www.fanfiction.net/https://www.fictionpress.com/https://manybooks. net/ Scope: Datasets examples that will benefit the eng team if possible to use: Author: Alex Yu Date: 4/1/2023 2:17:00 AM Range: Thank you, @meta.com Scope: Datasets examples that will benefit the eng team if possible to use: Author: Melanie Kambadur Date: 4/5/2023 9:52:00 PM Range: I suppose the framing of this doc limits to books/long-form writing, Redacted - Privilege Redacted - Privilege Scope: Author: Alex Yu Date: 4/14/2023 9:41:00 PM Range: Are these launch blocking? How high is the priority? Scope: Author: Alex Yu Date: 4/1/2023 12:51:00 AM meta.com , please review this section. Assigned to Range: @meta.com

Scope: Appendix: